Estableciendo las Bases para el Análisis de Datos en la Lotería Chaqueña

Conference Paper · November 2018

CITATIONS
0

READS
1,607

3 authors:

Patricia Andrea Loto
National University of the Northeast
2 PUBLICATIONS **0** CITATIONS
SEE PROFILE

Laura Cunico
Ingeniería Argentina (INGAR-UTN-CONICET), Argentina, Santa Fe
15 PUBLICATIONS **31** CITATIONS
SEE PROFILE

Melina Vidoni
Australian National University
54 PUBLICATIONS **126** CITATIONS
SEE PROFILE

Estableciendo las Bases para el Análisis de Datos en la Lotería Chaqueña

Patricia Loto[1], Laura Cunico[2], Melina Vidoni[2]
[1]*Lotería Chaqueña, patricialoto@hotmail.com*
[2]*INGAR CONICET-UTN, {laura-cunico, melinavidoni}@santafe-conicet.gov.ar*

Resumen

La tecnología y hardware actualmente permiten que las organizaciones generen una gran cantidad de datos en cada una de sus operaciones. Mediante la aplicación de diversas técnicas, la ciencia de datos permite extraer de ellos información sustancial para el funcionamiento de las organizaciones. Para ello es necesario disponer de bases de datos consistentes que faciliten la construcción de herramientas de análisis. Esto ha motivado el desarrollo de esta propuesta, focalizada particularmente en el caso de la Lotería Chaqueña. El organismo cuenta con una vasta colección de registros, que se encuentran dispersos en múltiples bases. El objetivo de este artículo es presentar la problemática, detallar las técnicas empleadas para la limpieza y organización de datos, realizar un análisis inicial, y plantear la arquitectura de una herramienta que simplifique el análisis de datos y visualización de resultados a los tomadores de decisiones.

1. Introducción

La tecnología y hardware actualmente permiten que las organizaciones generen una gran cantidad de datos adicionales en cada operación [1]. Esto implica problemas tales como la limitación de acceso a los mismos [2], y la existencia de almacenamientos diversos que producen datos no estructurados, cambiantes y a menudo inconsistentes [3].

Esta problemática surge en la Lotería Chaqueña, un organismo público autárquico perteneciente al Gobierno de la Provincia del Chaco. Allí, se posee una elevada cantidad de datos disponibles sobre los sorteos, las apuestas y resultados del juego Quiniela, pero los mismos están dispersos, inconsistentemente persistidos, y no son utilizados de manera óptima para generar análisis para la toma de decisiones informada.

Usualmente, los datos obtenidos pueden explorarse para inferir nuevo conocimiento y realizar predicciones a futuro [4]. El análisis predictivo permite construir modelos para prever comportamientos futuros, buscando mejorar la toma de decisiones [5, 6]. No obstante, no es posible realizar esto sin llevar a cabo un análisis inicial de forma previa; el cual consiste en evaluar la calidad de los datos y establecer las preguntas de investigación [7]. Se realiza a continuación de la limpieza y previo a los estudios principales, y busca transformar los datos y definir métricas de calidad.

El entorno de programación y análisis estadístico R nuclea no sólo funcionalidades de limpieza, análisis y visualización de datos [8], sino que además ofrece capacidades para la extracción y transformación de los mismos [9], hasta la generación de herramientas de análisis interactivo [10]. Por este motivo, su uso en ciencia y análisis de datos ha crecido exponencialmente, logrando una amplia aceptación en el campo [6].

Algunos autores han analizado datos provenientes de loterías, pero con finalidades diferentes. Por ejemplo, un trabajo analiza, mediante inferencia estadística, si los números favorecidos son realmente aleatorios, evitando la existencia de patrones [11]. De forma similar, se realizó un estudio de aleatoriedad para la United Kingdom Lotto, separando los valores en dos períodos de tiempo [12]. Otros autores realizaron *surveys* sobre el impacto de la eficiencia de los mercados de loterías, y la existencia de sesgo en las apuestas [13].

Así, el presente trabajo se distingue de la literatura actual, ya que describe la organización y análisis inicial de los datos de la Lotería Chaqueña, utilizando R. Esto se realiza para sentar las bases que permitan definir algoritmos concretos de análisis de los datos, los cuales serán visualizados de forma interactiva en una herramienta online generada mediante R Shiny.

Este artículo se organiza de la siguiente forma. La Sección 2 presenta la Lotería Chaqueña, el problema de negocio, y los conceptos fundamentales sobre los que trabaja. Luego, la Sección 3 aborda las dificultades encontradas en el pre-procesamiento de datos y cómo se subsanaron, mientras que la Sección 4 describe la exploración inicial de los datos, definiendo preguntas de inves-

tigación que delinean los análisis a desarrollar. La Sección 5 presenta el diseño de la herramienta online a desarrollar. Finalmente, la Sección 6 concluye el artículo.

2. Problemática del Negocio

La Lotería Chaqueña es un organismo público autárquico de la Provincia del Chaco, que depende del Ministerio de Hacienda y Finanzas Públicas de la provincia. Su principal juego es la Quiniela, cuyas apuestas pueden ser hechas a la Quiniela Nacional (de la Ciudad de Buenos Aires), a la Provincia (de Buenos Aires), o a la Chaqueña.

No obstante, el organismo sólo sortea efectivamente la Quiniela Chaqueña, realizando la extracción mediante bolillero propio de los números favorecidos. Respecto a las dos primeras, sólo se limita a reproducir los números favorecidos que fueron obtenidos en las loterías correspondientes mediante bolillero externo. Si bien se dispone de datos de las tres modalidades, este trabajo se concentra en la Quiniela Chaqueña, iniciada en Junio 2009.

2.1. Estructura de la Quiniela Chaqueña

En forma resumida, cada sorteo implica la extracción de veinte números de cinco cifras (denominados *números favorecidos*) en horarios determinados. Allí, la última cifra de izquierda a derecha indica el orden del número favorecido dentro de la conformación del extracto

Los sorteos se programan de acuerdo a las fechas y horas establecidas con las otras loterías, con las cuales el organismo tiene convenios para realizar sorteos o utilizar sus extractos. Ejemplos de estos convenidos son la Lotería Nacional y Provincia de Buenos Aires.

Cada sorteo se asocia a un número de sorteo y un *tipo*. Éste último hace referencia al horario en el que se lleva a cabo el sorteo, y se representa en los datos como un *carácter*. Esta información puede encontrarse resumida en la Tabla 1.

Tabla 1. Tipos de sorteos de la Quiniela Chaqueña.

Carácter	Tipo	Horario
P	Primera	11:30hs
V	Matutina	14:00hs
T	Vespertina	17:30hs
N	Nocturna	21:00hs

Sin embargo, la premiación se realiza considerando la cantidad de cifras acertadas del número favorecido; además del total, éstas pueden ser la última, los dos últimas, o las tres últimas. Por ejemplo, para el número 4325, se podría apostar al 4325, al 325, al 25, o al 5. En caso que haya ceros a la izquierda, estos se consideran; por ejemplo, para el valor 6701, se puede apostar a uno (1) a una cifra, a 01 a las dos y así sucesivamente.

El monto a pagar por acierto depende de:

- El monto apostado por el jugador, siendo la apuesta mínima de un peso ($1).
- La cantidad de cifras acertadas.
- La posición elegida. Esta puede ser “a la cabeza”, apostando al primer premio, “a las cinco” o “a las diez”, apostando a las primeras cinco o diez posiciones, respectivamente.
- Un coeficiente definido por la Lotería Chaqueña: 5, 70, 500, 3500; según la cantidad de cifras apostadas, y dividido por el alcance de premio al que se realizó la apuesta.

Por lo tanto, a mayor cantidad de cifras apostadas y cuanto más cerca esté el orden elegido del primer lugar, mayor es el monto a pagar por la Lotería. Por ejemplo, si la apuesta fuese de un peso, los aciertos se abonarían siguiendo el sistema de apuestas definido por la estructura de la Tabla 2. Así, si se acierta la última cifra “a la cabeza”, se paga cinco veces lo apostado, pero si se acierta a las dos últimas cifras a la misma posición se paga setenta veces lo apostado, y así sucesivamente.

Tabla 2. Estructura de coeficientes según apuestas y aciertos.

Cifras				
Cuatro	**Tres**	**Dos**	**Una**	**Apuesta $1**
3500	500	70	5	A la Cabeza
700	100	14	1	A los Cinco
233,33	50	7	0.5	A los Diez

Finalmente, la ganancia neta de cada sorteo se calcula usando la fórmula (E1), donde la *recaudación* es la suma de los montos de cada apuesta neta (no anulada) realizada para un sorteo específico, los *aciertos* corresponden a la suma de los montos a pagar por cada apuesta neta premiada, y la *comisión* es equivalente al veinte por ciento (20%) sobre el monto de la recaudación por las apuestas netas pagado a las Agencias oficiales por el cumplimiento del ciclo completo del Juego.

$$neto = recaudación - (aciertos + comisión) \quad \text{(E1)}$$

Debido a este esquema, resulta posible que ante un cierto sorteo, el neto resulte negativo y sea necesario pagar más de lo que se recaudó, incurriendo en una pérdida para el organismo

Existe un límite fijado por el organismo, denominado *Tope de Banca*, el cual se define como la suma equivalente a dos (2) veces la recaudación total para un sorteo específico. Además, se denomina *Salto de Banca* al evento en el cual el monto de premios a pagar supera al “tope de banca” para la jugada.

Estas situaciones hacen que la organización incurra en un riesgo que, debido a la situación actual de los datos, no es posible evaluar. La iniciativa de realizar un análisis de riesgo para descubrir si existen condiciones que ase-

guren un "salto de banca", surge como una alternativa promisoria para desarrollar capacidad de predicción y estrategias para la absorción del riesgo.

3. Pre-Procesamiento de Datos

La Lotería Chaqueña cuenta actualmente con una aplicación desarrollada en Foxpro 2.6. Dicha herramienta fue muy útil como primera aplicación, pero se encuentra en desventaja debido a la falta de robustez, escalabilidad y seguridad en comparación a las aplicaciones web modernas. Esto llevó al organismo a formar un pequeño equipo dedicado a la migración y actualización de sus sistemas informáticos. Sin embargo, dicho proceso no es tan rápido como el avance de la tecnología ni como la demanda de nuevos requerimientos por parte de los usuarios finales. Por esto mismo, aún no alcanza a las estructuras de datos empleadas en este trabajo.

Debido a estas limitaciones, la Lotería Chaqueña sólo mantiene en la base principal la información del año en curso. Los años anteriores son extraídos y almacenados como archivos comprimidos. Esto compromete la disponibilidad de los datos.

3.1. Proceso de Extracción

En las estructuras generadas, los datos se encuentran aislados, por lo que no resulta posible unificarlos en la consulta de extracción. Estas estructuras se organizan en:

- Estadística: ordenada por sorteo, tipo y fecha. Aquí se detalla la recaudación, el monto de aciertos y la cantidad de apuestas. Estos valores son totales del sorteo, y no se distinguen por jugada individual, ni agencia.
- Extractos: datos de sorteos de las tres modalidades (Nacional, Provincia y Chaqueña). Cada registro contiene números favorecidos, identificadores del personal involucrado, coeficientes de pago, número de acta, entre otros. Para los sorteos Nacionales y Provinciales, se detalla la recaudación por agencias provinciales y locales, según corresponda.
- Apuestas Individuales: Registro de la secuencia de números elegida por cada jugador, con fecha y hora de la apuesta, agencia, importe, entre otros; existiendo un archivo de apuestas por cada sorteo. Dado que los tickets son al portador, no se guarda información que identifique a la persona apostadora de forma individual.

Dichas estructuras se almacenan en tablas con formato DBF. Hay que destacar que estos archivos, son archivos de exportación de FoxPro, los cuales no pueden consultarse como una base de datos tradicional [14]. Si bien pueden accederse mediante sentencias propias, presentan problemas de consistencia e integridad de datos.

A su vez, existen otras estructuras con información circundante que no atañen a datos específicos de jugadas, y que no son consideradas para este análisis. Un ejemplo de esto son datos de las Agencias, como: titular, zona, dirección, código, código postal, entre otros.

De forma inicial, y para mantenerse en el marco de la Ley 25.326 de protección de datos personales [15], sólo se trabaja con las tablas Estadística y Extractos. A su vez, se selecciona el rango de datos desde el inicio de la Quiniela Chaqueña en Junio del 2009, hasta Diciembre de 2017, omitiendo los valores para el año en curso a fin de mantener una consistencia entre los períodos de estudio considerados.

Como se tratan de años previos, las tablas elegidas se encuentran almacenadas como archivos DBF. El procesamiento de estos datos se resume en la Figura 1.

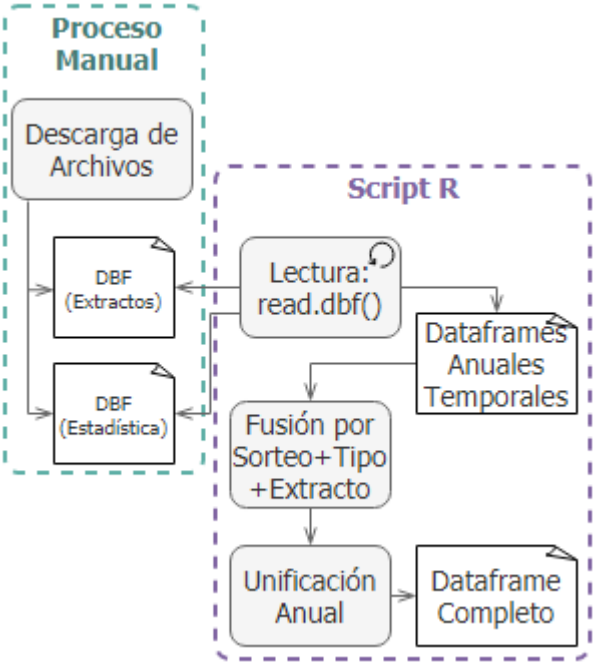

Figura 1. Proceso de extracción y conversión inicial de datos.

De esta forma, se ha logrado obtener una estructura de datos ordenada, a partir de la cual es posible realizar una limpieza y transformación, con el objetivo de iniciar el análisis de los datos.

3.2. Filtrado y Limpieza de Datos

Inicialmente, en la tabla Extracto, los datos de la lotería Nacional, Provincia de Bs. As. y Chaqueña se encuentran unificados. Así, el *data-frame* importado (ver Figura 1) mantiene esta estructura. Surgen dos problemas:

- Se generan columnas adicionales (o atributos) que son *equivalentes*. Por ejemplo, hay columnas con números favorecidos a nivel Nacional, Provincia, y Chaqueña, y montos a pagar en los tres niveles.
- Debido a la estructura de la anterior base de datos,

existen columnas denominadas *irrelevantes*: su valor es siempre el mismo para todos los registros. Esto se debe a que dichos campos dejaron de registrarse, pero las columnas no se eliminaron.

En consecuencia, la organización y limpieza de los datos aplica los pasos visibles en la Figura 2.

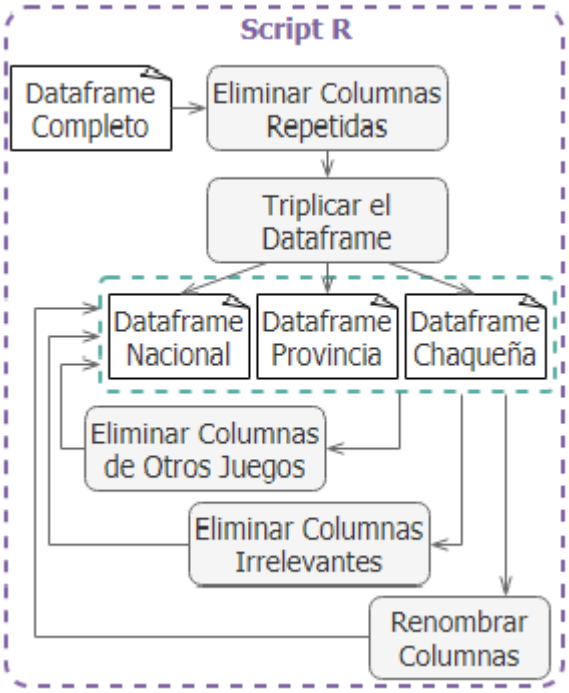

Figura 2. Proceso de organización de los datos.

Primero, se decide separar los datos en diferentes *data-frames*, y nombrar a las columnas equivalentes con los mismos nombres. Segundo, para cada *data-frame* individual se eliminan las columnas pertenecientes a otras loterías, luego las columnas irrelevantes, y finalmente se renombran para que la estructura sea equivalente. El objetivo de esto es lograr estructuras similares a las que después se les pueda aplicar sistemáticamente el mismo script de procesamiento.

En el segundo paso, de un total de 121 columnas, 75 eran equivalentes en los tres juegos, 27 eran irrelevantes, y 3 repetidas al fusionar las tablas (según Figura 1). Es importante destacar que muchas de las irrelevantes, eran también equivalentes. Debido a la cantidad de columnas afectadas, la depuración de las mismas se vuelve un punto vital para la mejora de la calidad de los datos.

Para el caso de Quiniela Chaqueña, se obtuvo una estructura de 31 columnas, y más de 10000 registros, desde Junio 2009 hasta Diciembre 2017, ambos inclusive.

3.3. Actualización de Montos

En todas las tablas, los montos de dinero se encuentran registrados en el valor en pesos argentinos ($ARS) del momento en que se llevó a cabo el registro. Esto origina inconsistencias en el tiempo, ya que no consideran la devaluación de la moneda; por ejemplo, $10.000 pesos en 2009, no resultan equivalentes a $10.000 en 2017. Esta problemática puede observarse en la Figura 3, donde se visibiliza un crecimiento exponencial en las recaudaciones y aciertos.

Figura 3. Recaudaciones y aciertos para la Lotería Chaqueña, en el período 2009-2017, en pesos argentinos.

Esto puede ser resuelto de dos formas. Por un lado, puede realizarse una deflactación o actualización del monto, usando como referencia al IPC, Índice de Precios al Consumidor, histórico [16], que implica llevar el monto a valores pasados o futuros. Otra alternativa, la adoptada en este trabajo, consiste en dolarizar los montos de dinero. Si bien el dólar estadounidense (U$D) también tiene variaciones, las mismas son despreciables para el tipo de análisis que se busca realizar. Más aún, la cotización del dólar en el período 2009-2017 [17] (ver Figura 4) presenta un crecimiento exponencial.

Figura 4. Cotización histórica del peso argentino en dólar.

Es por esto que, para trabajar con los montos de dinero de forma consistente, se dolarizan los registros empleando el valor más cercano. Para esto, se toman dos cotizaciones por mes: a principios (día 1) y a mediados (día 15).

La Figura 5 muestra la conversión de los datos presentados en la Figura 3. El comportamiento de la curva cambia radicalmente tras la dolarización.

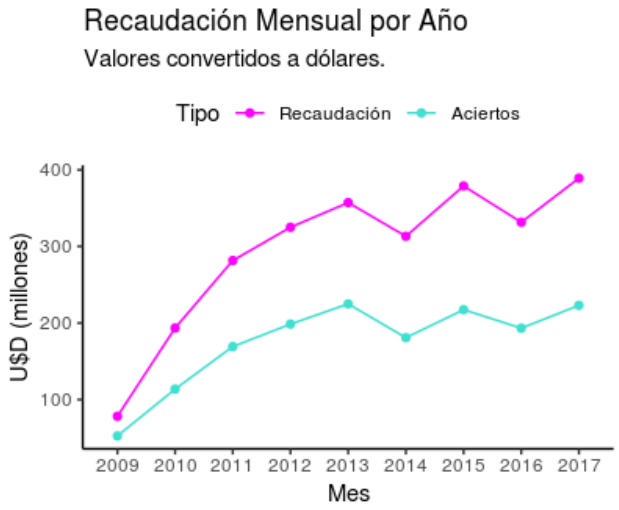

Figura 5. Recaudaciones y aciertos para la Lotería Chaqueña, en el período 2009-2017, en dólares.

4. Análisis Inicial

Tras realizar los procesos de extracción, unificación, limpieza y actualización, es posible realizar un estudio inicial para establecer lineamientos de análisis, dirigidos por preguntas de investigación. En una primera instancia, surgen tres posibles análisis.

4.1. Tendencia y Series Temporales

Mediante un script en R, se concentran los datos por mes y por año, obteniendo valores anuales (discretizados mensualmente) de recaudaciones, pago de premios, netos y apuestas. A esto se adiciona un análisis por semestre y cuatrimestre orientados a detectar posibles comportamientos estacionales y tendencias.

Tabla 3. Resumen de recaudación cuatrimestral.

	Recaudación por Cuatrimestre (U$D)		
Año	**Primero**	**Segundo**	**Tercero**
2009		3,841,196.63	7,016,427.14
2010	7,621,483.92	9,305,900.32	10,389,516.52
2011	8,709,351.70	11,269,526.85	12,457,553.40
2012	10,875,923.46	13,839,850.54	13,047,799.09
2013	12,651,938.83	12,796,481.05	12,058,824.16
2014	9,378,386.96	10,705,984.29	12,596,929.82
2015	3,350,822.28	16,018,586.24	18,124,310.36
2016	1,877,830.31	15,058,543.75	14,935,757.12
2017	6,431,883.00	19,364,958.02	18,868,402.18

En particular, la Tabla 3 resume los valores de recaudación por cuatrimestre de forma anual. Allí se observa que la recaudación del segundo cuatrimestre supera a la del primero en todos los casos; además, la recaudación del tercer cuatrimestre es mayor que la del primer cuatrimestre del año siguiente en la mayoría de los casos.

Cabe destacar, que en la Figura 6 se visualizan los montos de recaudación, convertidos a dólares, donde cada línea representa un año diferente. A excepción del año 2009, para el cual no se dispone de los datos del primer semestre, se observa una tendencia de crecimiento entre los meses de Julio y Septiembre, que posteriormente decae.

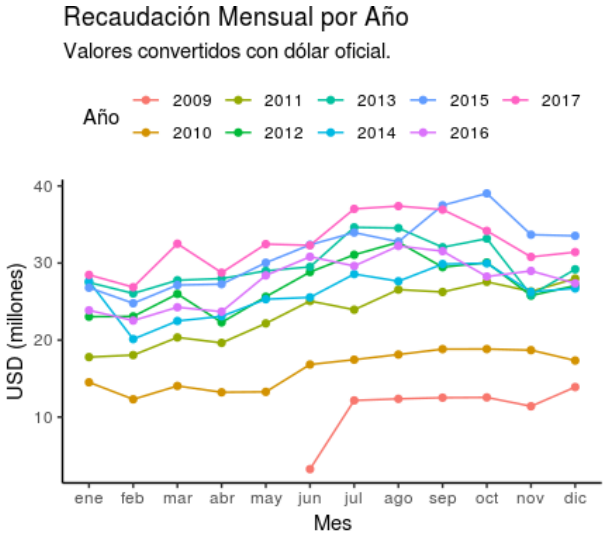

Figura 6. Recaudación mensual por año de la Quiniela Chaqueña.

Hay que destacar que el total de apuestas mensuales presenta un patrón similar. No obstante, el monto mínimo por jugada no ha cambiado acorde al IPC, determinando entonces que en los últimos años se ha producido un incremento de las mismas. Desde esta primera perspectiva surgen tres preguntas de investigación (PI):

- **PI-1.** ¿Existe una estacionalidad en los datos (recaudación y cantidad de jugadas), la cual corrobore el patrón observado?
- **PI-2.** ¿Es posible utilizar estos valores para predecir montos de recaudación para lo que resta de 2018 y para 2019?
- **PI-3.** ¿Por qué a partir del año 2015 la recaudación en el primer cuatrimestre decae tanto en comparación con la recaudación histórica del segundo y tercer cuatrimestre?

En ambos casos, se plantea la posibilidad de aplicar análisis de *series temporales* para responder dichas hipótesis [18]; [19]. Estas son secuencias de datos ordenados de forma temporal, tomados en puntos equidistantes; este algoritmo es muy utilizado en estadística, reconocimiento de patrones, predicciones y matemática financiera [20].

Así, con los datos disponibles es posible comparar los resultados con tendencias macro-económicas o de coyuntura para la Argentina, comportamientos económicos

sociales –por ejemplo, fechas turísticas, eventos sociales que impliquen gastos adicionales, etc.- entre otros, con el fin de discernir posibles motivos para tales movimientos.

Sin embargo, hay que destacar que no se consideran los montos a pagar en premios, ya que los mismos se determinan conforme a la cantidad de aciertos.

4.2. Agrupamiento de Condiciones

Un punto importante a analizar son las condiciones de pérdida. Si bien el organismo sólo registra Ganancia, Pérdida y "salto de banca", se decide agregar una condición de ganancia adicional, comparable a éste último. Esta *ganancia extrema* corresponde a los casos en los cuales la recaudación es al menos cuatro veces mayor que el monto a pagar en premios. Para el total de registros, la Tabla 4 resume los resultados para la Quiniela Chaqueña.

Tabla 4. Proporción de resultados para la Quiniela Chaqueña, respecto del total de sorteos (2009-2017).

Resultado	Cantidad	Porcentaje
Ganancia Extrema	3372	33%
Ganancia	5404	53%
Pérdida	916	9%
Salto de Banca	508	5%

Así, se observa que la mayoría de los casos representan condiciones de ganancia. Sin embargo, si se enfoca el análisis en las *excepciones*, es decir "saltos de banca" y "ganancia extrema", se puede plantear la posibilidad de relacionar esto con la cantidad de apuestas. La Tabla 5 suma estos datos por cuatrimestre.

Tabla 5. Cantidad de resultados extremos por cuatrimestre, para la Quiniela Chaqueña.

Cuatrimestre	Resultado	
	Salto de Banca	G. Extrema
Enero – Abril	149	1035
Mayo – Agosto	169	1127
Septiembre - Diciembre	190	1210

No obstante, si se observan estas cantidades para cada año, el comportamiento individual varía. La Figura 7 muestra, año a año, la evolución de los "saltos de banca". Se observa allí que hubo más ocurrencias en 2012, mientras que desde 2015 en adelante, la tendencia es de decrecimiento.

Teniendo en cuenta esto, surgen otros lineamientos de investigación:

- **PI-4.** ¿Se puede afirmar que existe una relación entre los resultados excepcionales, la cantidad de apuestas y los meses del año?
- **PI-5.** ¿Es posible modificar los coeficientes de pago según los meses, para disminuir los resultados que presentan "salto de banca"?
- **PI-6.** ¿Es posible determinar la probabilidad de que ocurra una excepción, dado que sucedió otra, en un período de tiempo (mismo día o semana, mismo mes del año actual o anterior)?

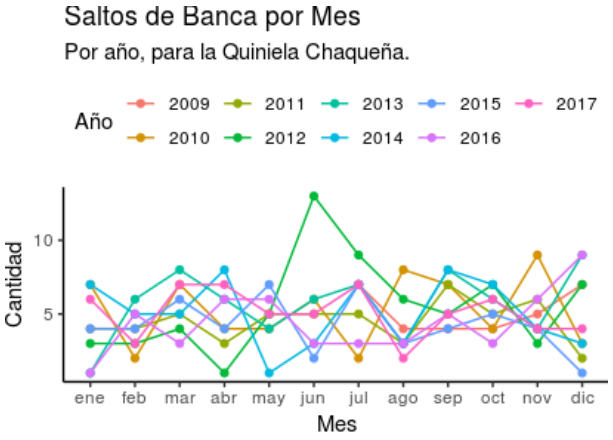

Figura 7. "Saltos de banca" mensuales anuales (2009 a 2017).

Para resolver las PI 4 y 5, se propone realizar análisis de clustering [21] [22]. Esta es una técnica de aprendizaje no supervisado que permite agrupar objetos de forma tal que los que pertenezcan a un mismo conjunto, tengan mayor similitud en sus atributos [23]. Así, es posible encontrar características comunes en estos casos, usando esta información para predecir comportamiento en caso de modificar los coeficientes.

Por otro lado, la PI6 puede estudiarse aplicando Naive-Bayes, la cual implica un modelo de probabilidad condicionada, es decir, descubrir las posibilidades de un evento, dado que ocurrió otro [24]. Otros autores como [25] ya han recurrido a esta técnica para realizar análisis semejantes, Probabilidad de Pérdida o Ganancia.

Dado que los aciertos se consideran por cifras (como se detalló en la Sección 2.1), del total de más de 206000 números favorecidos, se obtienen se obtienen más de 706000 posibles aciertos. Como todo número siempre se divide en al menos un dígito, éstos son los que tienen más apariciones.

Tabla 6. Cantidad de resultados excepcionales para los números favorecidos más frecuentes.

Número	Apariciones	Resultado	
		G. Extrema	Salto
8	20615	0.529	0.049
6	20593	0.532	0.049
94	2165	0.530	0.052
81	2156	0.510	0.047
549	252	0.541	0.041
365	250	0.534	0.033
2219	39	0.605	0.026
7248	38	0.448	0.079

La Tabla 6 presenta los dos valores más recurrentes para cada cantidad de cifras, con respecto a la proporción

de “ganancias extremas” y los “saltos de banca”. Si bien puede existir una similitud en las mismas, ésta debe ser comparada a la media y a la existente en números de menor frecuencia, para determinar si se trata de casos excepcionales.

Considerando que para cada sorteo se extraen veinte números favorecidos, es posible ampliar la dirección de investigación:

- **PI-7.** ¿Existe alguna relación de probabilidad entre los números favorecidos y los resultados obtenidos?
- **PI-8.** ¿Hay algún número que tenga mayor probabilidad de “ganancia extrema” o “salto de banca”?

5. Observatorio Quiniela: Requerimientos y Arquitectura

El interés de realizar estudios predictivos o de inferencia sobre los datos surge de la necesidad de Lotería Chaqueña de minimizar el riesgo de pérdida, al mismo tiempo que se aumenta su recaudación. El objetivo de dicho organismo público es el logro de fondos para contribuir a la política de acción social fijada por el Gobierno Provincial. Así, el 92% de la recaudación total se destina a dicho fin, quedando sólo el 8% para la operatoria administrativa.

Por esto, la plataforma de análisis a desarrollarse, denominada *Observatorio-Quiniela* será de suma importancia debido a que contribuirá al alcance de los objetivos del ente provincial de juegos de azar. Dicha plataforma será de uso interno, dirigida a la alta gerencia y tomadores de decisiones como stakeholders principales. En consecuencia, la capacidad de visualización, interacción y simplicidad de uso se vuelven aspectos fundamentales.

Dentro de R, existe el paquete Shiny [10], el cual permite crear aplicaciones online con estas mismas cualidades. Muchos autores lo han utilizado en diversos campos. Williams [26] desarrolló una aplicación de tipo educativo para mejorar el aprendizaje del concepto de intervalos de confianza; la misma tuvo resultados positivos, mejorando la atención de los alumnos. Por su parte, Matías y colaboradores [27] generaron una aplicación Shiny para ejecutar análisis biométricos y estadísticos usando modelos mixtos y multi-variados respecto al cultivo selectivo de plantas. Otra aplicación destacable es Ioncopy, que facilita el análisis de aberraciones detectables en genomas tumorales; la misma ha sido validada y utilizada de forma favorable [28].

A su vez, se plantea un desarrollo ágil, iterativo e incremental, donde en cada iteración se desarrolle un módulo que resuelva una (o un conjunto) de PI. De esta forma, la arquitectura general (ver Figura 8) será refinada en cada iteración, hasta obtener la estructura completa.

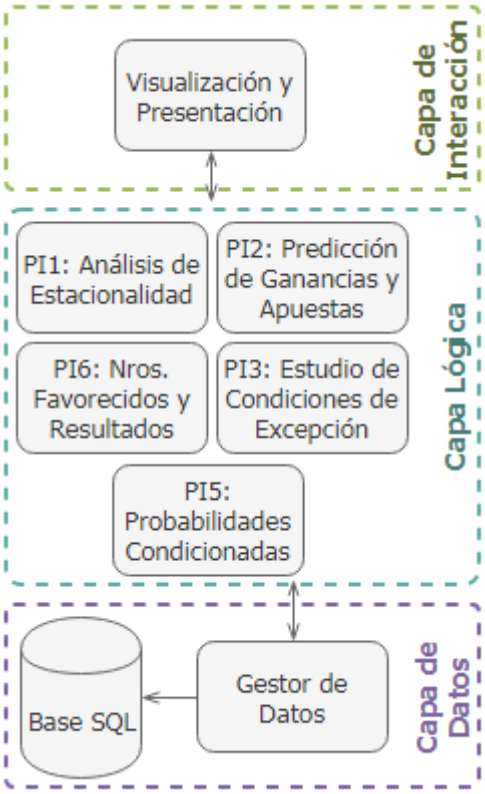

Figura 8. Modelo general para la arquitectura de la aplicación Observatorio-Quiniela.

A través del análisis inicial realizado en este artículo, se destaca que los datos, una vez organizados, respetarán una estructura definida. Diversos estudios han demostrado que la utilización de bases relacionales en ciencia de datos es positiva en la presencia de datos estructurados [29]. A su vez, como se destacó al inicio de la Sección 3, el organismo está llevando a cabo una migración de sus sistemas, hacia nuevas tecnologías.

Por esto mismo, se decide utilizar este almacenamiento para persistir los datos empleados. De esta forma, el proceso de organización discutido en la Sección 3 desemboca en *data-frames* almacenados en memoria. Éstos son convertidos a una base SQL Server, a través de un proceso de tres pasos: (1) selección del *data-frame*, (2) generación de la tabla y (3) carga de datos.

Hay que destacar que no todos los *data-frames* son persistidos: aquellos generados mediante la manipulación de datos base, no lo son. Un ejemplo de esto son las agregaciones mensuales que se realizan por año.

6. Conclusiones

Actualmente las organizaciones son capaces de reunir una importante cantidad de datos que contienen información sustancial tanto a nivel operativo cotidiano, como para su proyección a futuro. En el caso de la Lotería

Chaqueña, la capacidad de almacenamiento de registros de apuestas y resultados dota al organismo de información relevante para la predicción de tendencias de juego y la reducción del riesgo de pérdida. Sin embargo, estos valores se encuentran dispersos y mal documentados, lo cual dificulta la extracción de información de interés, y su posterior análisis para la toma de decisiones estratégicas fundada en datos.

La presente propuesta corresponde a la fase inicial de un proyecto de desarrollo de una herramienta online interactiva para el análisis y control de riesgo de pérdida de la Lotería Chaqueña, generada en R Shiny. La motivación principal radica en, mediante la explotación de los datos almacenados, en disponer de información precisa, confiable y actual sobre las tendencias de los distintos parámetros involucrados, como recaudación, premios a pagar, entre otros. De esta manera, sugerir mejoras en las reglas de negocio que permitan disminuir al mínimo el riesgo de pérdidas, sobre todo de saltos de banca; a su vez, permite tomar medidas para incrementar lo recaudado, y los aportes que vuelven a la ciudadanía mediante la asistencia social que realiza el organismo.

El entorno de programación escogido, R, está provisto de las funcionalidades de limpieza, extracción, transformación, análisis y visualización de datos, que lo posicionan como una alternativa conveniente para los fines del proyecto.

Específicamente, en este trabajo se describe el funcionamiento del organismo y la problemática detectada. Se detallan las técnicas implementadas para la organización, limpieza y análisis inicial de los registros existentes. Se exponen algunos lineamientos de análisis y definición de algoritmos concretos de estudio de los datos. Además, se plantea la arquitectura de la herramienta a desarrollar, con el requisito de simplificar el análisis y visualización de resultados a los tomadores de decisiones.

Cabe destacar que como resumen de la etapa inicial, el alcance de este trabajo no es exhaustivo y está sujeto a modificaciones futuras.

7. Referencias

[1] X. Wu, X. Zhu, G.-Q. W y W. Ding, «Data Mining with Big Data,» *IEEE Transactions on Knowledge and Data Engineering*, vol. 26, nº 1, pp. 97-107, 2014.

[2] G. Cimini, T. Squartini, D. Garlaschelli y A. Gabrielli, «Systemic Risk Analysis on Reconstructed Economic and Financial Networks,» *Scientific Reports*, vol. 5, pp. 15758:1-15758-13, 2015.

[3] B. Ayyub, Risk Analysis in Engineering and Economics, Segunda ed., Nueva York, USA: Taylor & Francis Group.

[4] A. Gandomi y M. Haider, «Beyond the Hype: Big Data Concepts, Methods, and Analytics,» *International Journal of Information Management*, vol. 35, nº 2, pp. 137-144, 2015.

[5] S. Finlay, Predictive Analytics, Data Mining and Big Data: Myths, Misconceptions and Methods, Primera ed., Reino Unido: Palgrave McMillan, 2014.

[6] G. James, D. Witten, T. Hastie y R. Tibshirani, An Introduction to Statistical Learning: with Applications in R, Springer Texts in Statistics ed., Nueva York, USA: Springer, 2013.

[7] H. Adèr, «Chapter 14: Phases and initial steps in data analysis,» de *Advising on Research Methods: A Consultant's Companion*, Primera ed., vol. 1, G. Mellenbergh, H. Adèr y D. Hand, Edits., Huizen, Johannes van Kessel Pub., 2008, pp. 333-356.

[8] R. Ihaka y R. Gentleman, «R: a language for data analysis and graphics,» *Journal of computational and graphical statistics*, vol. 5, nº 3, pp. 299-314, 1996.

[9] R. Peng, R Programming for Data Science, Primera ed., USA: Lulu.com, 2016.

[10] Studio Inc., «shiny: Web Application Framework for R,» 2017. [En línea]. Available: https://cran.r-project.org/web/packages/shiny/index.html. [Último acceso: 2018].

[11] J. Haigh, «The Statistics of Lotteries,» de *Handbook of Sports and Lottery Markets*, Elsevier, 2008, pp. 81-502.

[12] H. Okagbue, M. Adamu, P. Oguntunde, A. Opanuga y M. Rastogi, «Exploration of UK Lotto results classified into two periods,» *Data in Brief*, vol. 14, pp. 213-219, 2017,.

[13] W. Ziemba, «Efficiency of Racing, Sports, and Lottery Betting Markets,» de *Handbook of Sports and Lottery Markets*, D. Hausch y W. Ziemba, Edits., Elsevier, 2008, pp. 183-222.

[14] A. Simpson, Understanding dBASE III Plus Academic Edition, Primera ed., Alameda, USA: SYBEX Inc., 1990.

[15] Senado y Cámara de Diputados de la Nación Argentina, *Protección de los Datos Personales*, Capital Federal, Argentina: Boletín Oficial dela República Argentina, 2000.

[16] G. Milanesi, «Inflación y Descuento de Flujos de Fondos en Dos Monedas. Un Enfoque Integral,» *Revista Argentina de Investigación en Negocios (RAIN)*, vol. 3, nº 1, pp. 89-108, 2017.

[17] Banco Nación, «Cotizaciones Históricas del Dólar y Euro,» 2018. [En línea]. Available: http://www.bna.com.ar/Personas. [Último acceso: 2018].

[18] M. I. Landaluce Calvo, «Análisis exploratorio de estructuras temporales,» *Revista de Métodos Cuantitativo para la Economía y la Empresa*, pp. 55-77, Diciembre 2016.

[19] L. A. &. F. López-Rodríguez, «Discriminación temporal condicionada: una evaluación de efectos de,» *Journal of Behavior, Health & Social Issues*, pp. 62-69, 2017.

[20] H. Kantz y T. Schreiber, Nonlinear Time Series Analysis, Segunda ed., Cambridge: Cambridge University Press, 2004.

[21] G. Tardiolia, R. Kerrigana, M. Oates, J. O'Donnell y D. P. Finn, «Identification of representative buildings and

building groups in urban datasets using a novel pre-processing, classification, clustering and predictive modelling approach,» *Building and Environment*, pp. 90-106, 2018.

[22] C. Martellaa, A. Miragliaa, J. Frosta, M. Cattanib y M. v. Steen, «Visualizing, clustering, and predicting the behavior of museum visitors,» *Pervasive and Mobile Computing*, pp. 430-443, 2017.

[23] D. Pfitzner, R. Leibbrandt y D. Powers, «Characterization and evaluation of similarity measures for pairs of clusterings,» *Knowledge and Information Systems*, vol. 19, p. 361–394, 2009.

[24] G. Webb, J. Boughton y Z. Wang, «Not So Naive Bayes: Aggregating One-Dependence Estimators,» *Machine Learning*, vol. 58, nº 1, p. 5–24, 2005.

[25] S. Maitraa, S. Madanb, R. Kandwalc y P. Mahajan, «Mining authentic student feedback for faculty using Naïve,» *Procedia Computer Science*, pp. 1171-1183, 2018.

[26] I. Williams y K. Williams, «Using an R shiny to enhance the learning experience of confidence intervals,» *Teaching Statistics*, vol. 40, nº 1, pp. 24-28, 2017.

[27] F. Matias, I. Granato y R. Fritsche-Neto, «Be-Breeder: an R/Shiny application for phenotypic data analyses in plant breeding,» *Crop Breeding and Applied Biotechnology*, vol. 18, nº 2, 2018.

[28] J. Budczies, N. Pfarr, E. Romanovsky, V. Endris, A. Stenzinger y C. Denkert, «Ioncopy: an R Shiny app to call copy number alterations in targeted NGS data,» *BMC Bioinformatics*, vol. 19, nº 157, 2018.

[29] F. Provost y T. Fawcett, «Data Science and its Relationship to Big Data and Data-Driven Decision Making,» *Big Data*, vol. 1, nº 1, pp. 51-59, 2013.